Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*

Athanasios Theologis[*,†], Joseph R. Ecker[*,‡,§], Curtis J. Palm[‖], Nancy A. Federspiel[§,‖], Samir Kaul[¶], Owen White[¶], Jose Alonso[‡], Hootan Altafi[‖], Rina Araujo[‖], Cheryl L. Bowman[¶], Shelise Y. Brooks[‡], Eugen Buehler[‡], April Chan[†], Qimin Chao[‡], Huaming Chen[‡], Rosa F. Cheuk[‡], Christina W. Chin[†], Mike K. Chung[†], Lane Conn [§,‖], Aaron B. Conway[‖], Andrew R. Conway[‖], Todd H. Creasy[¶], Ken Dewar[‡], Patrick Dunn[‡], Pelin Etgu[†], Tamara V. Feldblyum[¶], JiDong Feng[‡], Betty Fong[†], Claire Y. Fujii[¶], John E. Gill[¶], Andrew D. Goldsmith[†], Brian Haas[¶], Nancy F. Hansen[‖], Beth Hughes[†], Lucas Huizar[‖], Jonathan L. Hunter[‡], Jennifer Jenkins[¶], Chanda Johnson-Hopson[‡], Shehnaz Khan[‡], Elizabeth Khaykin[‡], Christopher J. Kim[‡], Hean L. Koo[¶], Irina Kremenetskaia[†], David B. Kurtz[‖], Andrea Kwan[†], Bao Lam[‖], Stephanie Langin-Hooper[‡], Andrew Lee[‡], Jeong M. Lee[†], Catherine A. Lenz[†], Joycelyn. H. Li[†], YaPing Li[‡], Xiaoying Lin[¶], Shirley X. Liu[†], Zhaoying A. Liu[†], Jason S. Luros[†], Rama Maiti[¶], Andre Marziali[§,‖], Jennifer Militscher[¶], Molly Miranda[‖], Michelle Nguyen[‖], William C. Nierman[¶], Brian I. Osborne[†], Grace Pai[¶], Jeremy Peterson[¶], Paul K. Pham[†], Michael Rizzo[¶], Timothy Rooney[¶], Don Rowley[‖], Hitomi Sakano[†], Steven L. Salzberg[¶], Jody R. Schwartz[†], Paul Shinn[‡], Audrey M. Southwick[‖], Hui Sun[‡], Luke J. Tallon[¶], Gabriel Tambunga[‡], Mitsue J. Toriumi[†], Christopher D. Town[¶], Teresa Utterback[¶], Susan Van Aken[¶], Maria Vaysberg[†], Valentina S. Vysotskaia[†,§], Michelle Walker[‡], Dongying Wu[¶], Guixia Yu[†], Claire M. Fraser[¶], J. Craig Venter[§,¶] & Ronald W. Davis[‖]

[†] Plant Gene Expression Center/USDA-U.C. Berkeley, 800 Buchanan Street, Albany, California 94710, USA
[‡] Plant Science Institute, Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA
[‖] Stanford Genome Technology Center, 855 California Avenue, Palo Alto, California 94304, USA
[¶] The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, Maryland 20850, USA
[*] These authors contributed equally to this work

The genome of the flowering plant *Arabidopsis thaliana* has five chromosomes[1,2]. Here we report the sequence of the largest, chromosome 1, in two contigs of around 14.2 and 14.6 megabases. The contigs extend from the telomeres to the centromeric borders, regions rich in transposons, retrotransposons and repetitive elements such as the 180-base-pair repeat. The chromosome represents 25% of the genome and contains about 6,850 open reading frames, 236 transfer RNAs (tRNAs) and 12 small nuclear RNAs. There are two clusters of tRNA genes at different places on the chromosome. One consists of 27 tRNA^{Pro} genes and the other contains 27 tandem repeats of tRNA^{Tyr}-tRNA^{Tyr}-tRNA^{Ser}genes. Chromosome 1 contains about 300 gene families with clustered duplications. There are also many repeat elements, representing 8% of the sequence.

Future insights into basic plant biology and the ability to manipulate plants through genetic engineering for agronomic improvement will depend on whether we can identify the genes that control fundamental developmental and metabolic processes. The identification of these genes has until recently relied on their mutant phenotypes, genetic map positions and positional cloning approaches. Sequencing the complete genomes of model organisms[3–5] has greatly facilitated the identification of genes that are important for both basic biological and disease-related processes. Most plant species have relatively large genomes (>500 megabases (Mb)) with many repeated sequences. Complete genome sequencing efforts in plants have therefore centred largely on *Arabidopsis thaliana*[6,7], which has a small genome (around 130 Mb) with low repetitive DNA content[2,8,9]. The Arabidopsis Genome Initiative (AGI) was established in 1996 to sequence the genome of *Arabidopsis*. Here we report the DNA sequence of 28.8 Mb of chromosome 1 of ecotype Columbia, fully annotated into two contigs—the northern arm (approximately 14.2 Mb) and the southern arm (approximately 14.6 Mb).

Chromosome 1 is metacentric and was originally estimated, using the yeast artificial chromosome (YAC)-based physical map[10], to be 31-Mb long. Sequencing was initiated using bacterial artificial chromosomes (BACs) anchored to a physical map[11]. An optimal tiling path for sequencing was determined using a BAC end-sequencing strategy[12]. Three AGI groups, Genoscope (http:// www.genoscope.cns.fr/externe/arabidopsis), The Institute for Genomic Research (TIGR) (http://www.tigr.org/tdb/at/abe/bac_end_ search.html) and the Stanford/University of Pennsylvania/Plant Gene Expression Center (SPP) Consortium (http://genome.bio.upenn.edu/physical-mapping/bacendlist/bacends.html) determined the end sequences for almost all of the BACs in the TAMU[13] and

Table 1 Features of chromosome 1

Feature	Value
(a) The DNA molecule	
Length	28,762,046 bp
Top arm	14,172,442 bp
Bottom arm	14,589,604 bp
Base composition (% GC)	
Overall	35.8
Coding	43.8
Non-coding	31.8
Number of genes	6,848
Gene density	4.1 kb per gene
Average gene length	2,145 bp
Average peptide length	460 amino acids
Exons	
Number	35,768
Total length	9,396,363 bp (33%)
Average per gene	5.2
Average size	272 bp
Introns	
Number	28,951
Total length	4,807,531 bp (17%)
Average size	166
Number of genes with ESTs	3,448 (50%)
Number of ESTs* for chromosome 1 genes	~30,000 (27%)

(b) The proteome		
Classification/function	Number	Percentage (%)
Total proteins	6,848	100
With similarity to GenBank entries	4,793	70
Unknown	1,590	23
Hypothetical	1,935	28
With putative function	3,323	49
With putative signal peptides		
Secretory pathway	1,049	15
Chloroplast	521	7.5
Mitochondria	96	1.5
Classification of proteins with putative function		
Cellular metabolism	995	30
Transcription	498	15
Plant defence	354	11
Signalling	327	10
Growth and development	302	9
Protein fate	292	9
Intracellular transport	247	7
Ion transport	194	6
Protein synthesis	113	3
Total	3,322	100

* Around 110,000 ESTs used for analysis (http://www.arabidopsis.org).

[§] Present addresses: The Salk Institute for Biological Studies, 10010 N. Torrey Pines Road, La Jolla, California 92037, USA (J.R.E.); Exelixis, Inc., 170 Harborway, P.O. Box 511, South San Francisco, California 94083-0511, USA (N.A.F. and V.S.V.); DNA Sciences, Inc., 2375 Garcia Avenue, Mountain View, California 94043, USA (L.C.); Department of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, British Columbia V6T 1Z1, Canada (A.M.); Celera Genomics, Inc., 45 West Gude Drive, Rockville, Maryland 20850, USA (J.C.V.).

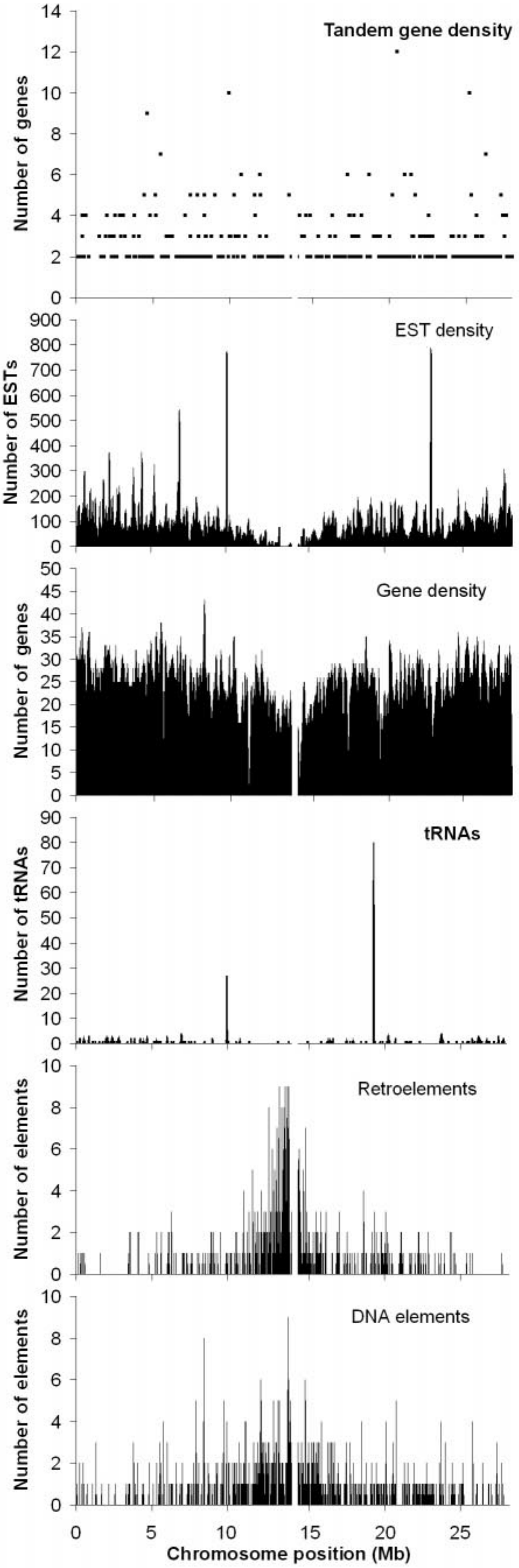

Figure 1 Density of various features along chromosome 1. The analyses were performed on the entire chromosome (28,762,046 bp) by joining the two arms at the centromeric gap, shown as a discontinuity. The densities of the features were obtained by analysing the sequence every 100 kb using a 10-kb sliding window. The analyses for retro- and DNA-transposable elements were carried out every 10 kb.

IGF[14] libraries, corresponding to about 36,000 unique end sequences from around 18,300 BAC clones (~14× coverage; 10,200 TAMU and 8,100 IGF clones). We searched completed BAC sequences against the end sequence database, and chose BAC clones with a minimum of sequence overlap for further sequencing. On average, we identified homologous BAC end sequences every five kilobases (kb) along the length of the query BAC sequence. We sequenced one YAC and 369 BAC clones to produce the finished unique sequence. Seventy-seven clones served as 'seed' points. The two contigs extend from the telomeres to the centromeric borders, which are rich in transposons, retrotransposons and diverse repetitive elements[15]. We identified the telomeric regions of the chromosome with the aid of previously isolated putative north (pAtT60) and south (ATYpATT1) telomeric clones from the Landsberg ecotype[16] as described (see Methods in the Supplementary Information). The GC content of the chromosome is low, which made it easier to sequence using the available chemistry (Table 1(a)). The northern arm does not have any sequencing gaps. The southern arm has three sequencing gaps at 15.349 Mb (BAC T18F15, C084807), 15.573 Mb (BAC T2P3, CO84820), and 20.070 Mb (BAC T4M14, AC027036).

We verified the collinearity of the assembled sequence using BLAST[17]analysis of the BAC end sequences against the assembled contig sequences. In addition, the location of 107 sequenced chromosome 1 markers on the assembled chromosome is collinear with their location on the recombinant inbred linkage map[18]. The sequence accuracy is estimated to be no more than one error in 25,000 bp, on the basis of the discrepancies found in the sequenced overlapping regions of BACs. The centromere is estimated to be 1.3-Mb long (refs 15, 19), bringing the overall size of the chromosome to about 30 Mb. The relationship between physical and genetic distance is around 230 kb cM^{-1}, uniform across the chromosome and similar to that reported for chromosomes 2 (ref. 6) and 4 (ref. 7).

We used gene prediction programs and database searches to annotate and determine some of the features of the chromosome. It contains about 6,850 putative genes at a density of 1 gene per 4.1 kb, which is more than twice the density found in *Drosophila*[5]. The average gene is about 2-kb long (from start to stop codon) and contains five exons with an average size of 272 base pairs (bp). Therefore, almost 50% of the chromosome consists of genes (Table 1(a)). Approximately 1,250 (18%) of the genes are intronless and many have been annotated as 'hypothetical' proteins. The largest gene, F5A8.4 (AC004146), contains 68 exons and encodes a 5,139-amino-acid polypeptide (relative molecular mass ~500,000 (M_r ~500K) which is a ubiquitous putative membrane protein found in *Saccharomyces cerevisiae*[3], *Caenorhabditis elegans*[4], *D. melanogaster*[5] and *Homo sapiens*. The smallest recognized gene, T6H22.15 (AC009894), encodes the L41 protein of the 60S ribosomal complex, which contains 25 amino acids. Expressed sequence tags (ESTs) exist for this gene. At least 3,448 (50%) genes are expressed, as shown by the presence of corresponding *Arabidopsis* complementary DNAs or ESTs in GenBank. Thirty-thousand (27%) of the 110,000 ESTs screened are products of chromosome 1 genes (Table 1(a)). Ninety-seven per cent of the chromosome 1 ESTs map to annotated genes. Genes that are most highly represented by ESTs include putative chlorophyll a/b-binding protein, small subunit of RuBP carboxylase, ferrodoxin/precursor and photosystem II 10K polypeptide. The gene density and EST distribution are more or less the same along the chromosome, with their lowest values around the regions flanking the centromere. There are a few highly dense EST regions with more than 700 ESTs per 100 kb, indicating highly transcribed regions (Fig. 1).

The chromosome contains 236 tRNAs, which is similar to the total number (293) found in the *Drosophila* genome[5]. The tRNAs are evenly distributed along the chromosome except for two regions where they cluster (Fig. 1). The first cluster (at 9.989 Mb) contains 27 tandem $\mathrm{tRNA}^{\mathrm{Pro}}$ genes, with the last gene being a pseudogene.

The other (at 19.282 Mb) has 27 tandem repeats of the tri-repeat tRNA^{Tyr}- tRNA^{Tyr}- tRNA^{Ser} (Fig. 2). The tRNA^{Ser}/tRNA^{Tyr} cluster has been structurally characterized[20] using conventional molecular approaches, but the tRNA^{Pro} cluster was previously undetected. The sequence homology among the tRNA repeats ranges from 98% to 99%. The three tRNA genes have the same orientation and are separated by 250 and 370 bp, respectively. It has been suggested that tRNA clustering may reflect tissue-specific co-regulation of the tRNA genes[20].

There are four pairs of megabase-scale segmental duplications within the chromosome. Their locations are: 3.1–3.8/19.3–19.8; 5.6–6/27.8–27.3; 6.2–7.5/25.4–26.7; and 8–8.6/25–24.3 Mb. The first and third duplications have the same orientation relative to the centromere; for the second and fourth set, the two members are in opposite orientations. There are also large-scale duplications between chromosome 1 and each of the other *Arabidopsis* chromosomes[19]. Analysis also reveals diverse repetitive elements representing 8% of the sequence (see Table 2 in the Supplementary Information) consisting of various retroelements (2.6%), DNA elements (2.4%) and a number of simple and low-complexity repeats (2.6%). There is an inverse relationship between gene density and retro- and DNA-element densities in the borders of the centromeric region (Fig. 1), a hallmark of such chromosomal domains[21].

Computational analysis predicts 6,848 proteins (Table 1(b)), corresponding to about a quarter of the *Arabidopsis* proteome and equal to half the predicted number of proteins in *Drosophila* (13,601 predicted proteins). Around one-third (28%) of the proteins are 'hypothetical', being predicted by various gene prediction programs, but without EST matches. Twenty-five per cent have an 'unknown' function, but we know that they are transcriptionally active because there is an EST. Seventy per cent of the annotated proteins have some similarity to other 'hypothetical', 'unknown' or 'putative function' proteins from plants and other eukaryotic genomes such as *S. cerevisiae*[3], *C. elegans*[4], *D. melanogaster*[5] and *H. sapiens*. Table 1(b) shows a functional classification of the proteins, based on their amino-acid motifs. Thirty per cent of the proteins with putative function participate in cellular metabolism, and another 50% are involved in transcription, plant defence, signalling, and growth and development. Comparison of the predicted proteins of chromosome 1 with other proteins from available complete genome sequences reveals that more than 1,500 proteins have significant homology (cutoff $\leq 10^{-30}$) to proteins from the available genomes (*S. cerevisiae*, *C. elegans*, *D. melanogaster* and *H. sapiens*). The top 14 homologies are shown in Table 3 of the Supplementary Information. Clearly, proteins involved in RNA splicing, tRNA biosynthesis, translation and metabolism are highly conserved throughout the eukaryotic kingdom.

The chromosome contains 312 families of tandemly arranged genes with 847 members (comprising 12% of all the genes that have a blast score of more than 200). There are between two and twelve members in each family (Fig. 3). These families are distributed throughout the chromosome (Fig. 1). One hundred and fifty-six families encode a set of distinct protein isoforms and are not members of superfamilies. The remaining 156 families are members of 46 superfamilies, which vary in size from 2 to 23 families. For example, the T28K15.6 (AC022522) superfamily contains five two-member families, which encode putative NBS/LRR disease-resistance proteins. Two of the T28K15.6 families are located at ~4 Mb and the other three at ~20.5 Mb on the chromosome. The families

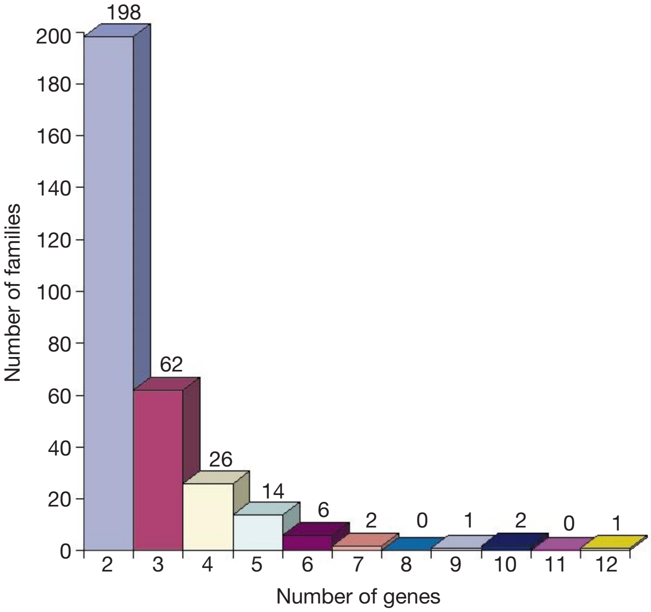

Figure 3 Frequency distribution of genes in multigene families with tandem gene arrangements.

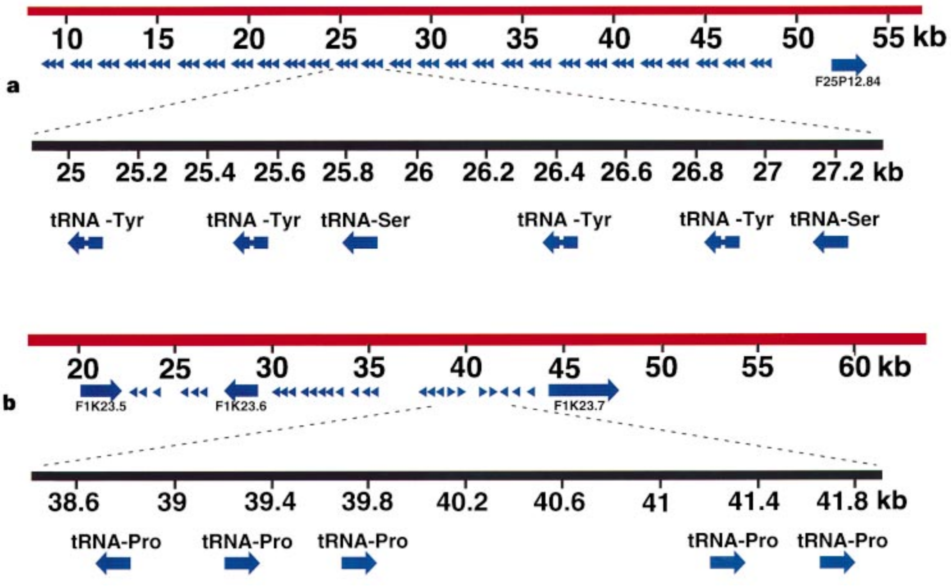

Figure 2 Clusters of tRNA genes in chromosome 1. **a,** The tRNA^{Ser}/tRNA^{Tyr} repeat cluster at 19.282 Mb is located on BAC F25P12 (accession no. AC009323). **b,** The tRNA^{Pro} cluster at 9.989 Mb is located on BAC F1K23 (accession no. AC007508).

of 25 superfamilies are close together (less than 2 Mb apart), whereas those of the remaining 21 are farther apart. In a few cases, the families are 20 Mb apart. For example, the T28P6.7 (AC007259) superfamily, which encodes S-locus-like receptor kinases (SRKs), has four families with two, three, twelve and two members located at 3.79, 38.1, 20.5 and 22.33 Mb, respectively. The first two SRK families are closely linked, separated by two unrelated genes. The SRK proteins have been implicated in controlling self-incompatibility in the reproductive system in *Brassicaceae*[22], and it has been proposed that the autogamous reproductive system of *Arabidopsis* reflects the absence of the SRK genes from the *Arabidopsis* genome[22]. It remains to be determined whether the 19 putative SRK isoforms encoded by the T28P6.7 superfamily and the 12 SRK isoforms encoded by the T28P6.2 superfamily (three families with three, seven and two members, respectively) on chromosome 1 have an S-locus protein kinase activity. The largest superfamily is F22L4.6 (AC061957) with 23 families, which encodes 78 isoforms of a putative protein kinase. A large percentage of the gene families on this chromosome encode proteins that are involved in disease resistance, cell wall degradation and secondary metabolite biosynthesis.

It will be important to find out the biological significance of the multigene families. Why do so many gene products encode isoforms of the same polypeptide? This question applies to gene families with tandem gene arrangement as well as those with dispersed gene arrangements, such as 1-aminocyclopropane-1-carboxylic acid (ACC) synthase (ACS). This family has two members (ACS2 and ACS10) on chromosome 1 and eight other members in the other four chromosomes. It has been suggested[23] that the presence of different *ACS* isoforms may reflect tissue-specific expression that satisfies the biochemical properties of the cells or tissues in which each is expressed. For example, if a group of cells or tissues have low concentrations of the substrate, S-adenosyl methionine, then these cells express a high affinity (low *K*m) *ACS* isoform. Accordingly, the distinct biological function of each isoform is defined by its biochemical properties, which in turn define its tissue-specific expression. Such a concept can accommodate gene families encoding structural proteins as well as enzymes.

The sequences of chromosome 1 and the other *Arabidopsis* chromosomes[6,7,24,25] will provide a wealth of information, but more experimental work is required to determine gene functions that will advance plant science[26]. Most of the genes and their predicted functions should be interpreted with caution. Mapping the transcriptional units of the *Arabidopsis* chromosomes in the future will provide experimental verification of the annotation. Chip technology[27] has the potential to facilitate the mapping of the transcriptional units, leading to the isolation of full-length cDNA clones for all the genes of the *Arabidopsis* chromosomes. The availability of the uni-cDNA collection will eliminate the need for cDNA library construction and will allow the creation of a library where every cDNA will be represented at equimolar concentration. More importantly, the full-length cDNA collection will allow the biochemical characterization of the *Arabidopsis* proteome. Concomitantly, the available DNA sequence and chip technology eliminate the need for Southern and northern analyses in *Arabidopsis*. Isolation of insertional mutants for all the genes on the chromosomes will offer additional resources to elucidate the function of the genes by reverse genetics. The biological resources that will be available for *Arabidopsis* will allow plant biologists to advance plant science and agriculture to levels never dreamed of 30 years ago. □

Methods

We used two BAC and one YAC libraries made from the ecotype Columbia of *Arabidopsis thaliana*: the TAMU[13] BAC library (T); the IGF[14] BAC library (F); and the yUP[28] YAC library. BAC DNA was hydrodynamically sheared[29] to 1–2 kb and the DNA was ligated into an M13 vector using a 'linker-adapter' system[30] that minimizes the formation of chimaeric DNAs. We used an automated sample preparation system for template preparation[31]. We sequenced each BAC to 10× coverage using Dyeprimer chemistry. Individual BACs were assembled from the shotgun sequences using Phred[32,33]/Phrap (http://www.phrap.org)/Consed9[34], and the gaps in each BAC were closed using a combination of BAC walking, directed PCR and resequencing of individual clones using Dyeterminator chemistry. All sequences are >97% double-stranded and have no more than one error per 25 kb. BACs were sequenced at TIGR as described[6].

We analysed the DNA sequence of each BAC for protein-coding genes using gene-prediction software and alignments with sequences from the EST and non-redundant protein databases. For initial gene and exon predictions, we used GRAIL[35], Genscan[36], fexa (V. Solovyev, http://genomic.sanger.ac.uk/gf/gf.shtml), GlimmerA[37] and GeneMark.HMM[38]. Splice site predictions were made using GeneSplicer (M. Pertea and S. Salzberg, unpublished software) and NetPlantGene[39]. We aligned sequences in the EST and non-redundant protein databases to the BAC sequences using the BLAST[17] and AAT[40] sequence database search and alignment software. Protein gene models were then assigned both automatically by the PEDANT[41] system and manually by evaluating and editing the gene prediction and database alignment data for each BAC. We identified tRNA genes using the program tRNA-scan-SE[42]. We identified potential transit and targeting peptide sequences in proteins using TargetP[43] (specificity >0.95), and repeats using Repeat Masker (A. F. Smit & P. Green; http://ftp.genome.washington.edu/RM/RepeatMasker.html; http:// www.girinst.org).

Received 20 October; accepted 15 November 2000.

1. Goodman, H., Ecker, J. R. & Dean, C. The genome of *Arabidopsis thaliana*. *Proc. Natl Acad. Sci. USA* **93,** 10831–10835 (1995).
2. Meyerowitz, E. M. in *Arabidopsis* (eds Meyerowitz, E. M. & Somerville, C.) 21–36 (Cold Spring Harbor Press, Cold Spring Harbor, NY, 1994).
3. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546–567 (1996).
4. *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282,** 2012–2046 (1998).
5. Adams, M. D. The genome sequence of *Drosophila melanogaster Science* **287,** 2185–2195 (2000).
6. Lin, X. *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402,** 761–768 (1999).
7. Mayer, K. Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402,** 769–777 (1999).
8. Mozo, T. *et al.* A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nature Genet.* **22,** 271–275 (1999).
9. Marra, M. *et al.* A map for sequence analysis of the *Arabidopsis thaliana* genome. *Nature Genet.* **22,** 265–270 (1999).
10. Creusot, F. *et al.* The CIC library: a large insert YAC library for genome mapping in *Arabidopsis thaliana*. *Plant J.* **8,** 763–770 (1995).
11. Ewens, W. J. *et al.* Genome mapping with anchored clones: theoretical aspects. *Genomics* **11,** 799–805 (1991).
12. Venter, J. C., Smith, H. O. & Hood, L. A new strategy for sequencing. *Nature* **381,** 364–366 (1996).
13. Choi, S., Creelman, R. A., Mullet, J. E. & Wing, R. Construction and characterization of a bacterial artificial chromosome library of *Arabidopsis thaliana*. *Plant Mol. Biol. Rep.* **13,** 124–128 (1995).
14. Mozo, T., Fischer, S., Shizuya, H. & Altmann, T. Construction and characterization of the IGF *Arabidopsis* BAC library. *Mol. Gen. Genet.* **258,** 562–570 (1998).
15. Round, E. K., Flowers, S. K. & Richards, E. J. *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. *Genome Res.* **7,** 1045–1053 (1997).
16. Richards, E. J., Chao, S., Vongs, A. & Yang, J. Characterization of *Arabidopsis thaliana* telomeres isolated in yeast. *Nucleic Acids Res.* **20,** 4039–4046 (1992).
17. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
18. Lister, C. & Dean, C. Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. *Plant J.* **4,** 745–750 (1993).
19. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000)..
20. Beier, D., Stange, N., Gross, H. J. & Beier, H. Nuclear tRNA(Tyr) genes are highly amplified at a single chromosomal site in the genome of *Arabidopsis thaliana*. *Mol. Gen. Genet.* **225,** 72–80 (1991).
21. Copenhaver, G. P. *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* **286,** 2468–2474 (1999).
22. Conner, J. A., Conner, P., Nasrallah, M. E. & Nasrallah, J. B. Comparative mapping of the Brassica S locus region and its homolog in *Arabidopsis*: implications for the evolution of mating systems in the *Brassicaceae*. *Plant Cell* **10,** 801–812 (1998).
23. Rottmann, W. E. *et al.* 1-aminocyclopropane-1-carboxylate synthase in tomato is encoded by a multigene family whose transcription is induced during fruit and floral senescence. *J. Mol. Biol.* **222,** 937–961 (1991).
24. Salanoubat, M. *et al.* Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*. *Nature* **408,** 820–822 (2000).
25. Tabata, S. *et al.* Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature* **408,** 823–826 (2000).
26. Chory, J. *et al.* National Science Foundation-sponsored workshop report: "The 2010 Project" functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. *Plant Physiol.* **123,** 423–426 (2000).
27. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405,** 827–836 (2000).
28. Ecker, J. R. PFGE and YAC analysis of the *Arabidopsis* genome. *Methods* **1,** 186–194 (1990).
29. Oefner, P. J. *et al.* Efficient random subcloning of DNA sheared in a recirculating point-sink flow system. *Nucleic Acids Res.* **24,** 3879–3886 (1996).
30. Dietrich, F. S. *et al.* The nucleotide sequence of *Saccharomyces cerevisiae* chromosome V. *Nature* (Suppl.) **387,** 78–81 (1997).
31. Marziali, A., Willis, T. D., Federspiel, N. A. & Davis, R. W. An automated sample preparation system for large-scale DNA sequencing. *Genome Res.* **9,** 457–462 (1999).
32. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using *Phred* I. Accuracy assessment. *Genome Res.* **8,** 175–185 (1998).
33. Ewing, B. & Green, P. Base-calling of automated sequencer traces using *Phred* II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
34. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res.* **8,**

195–202. (1998).

35. Uberbacher, E. C. & Mural, R. J. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl Acad. Sci. USA* **88,** 11261–11265 (1991).

36. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).

37. Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J. & Tettelin, H. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59,** 24–31 (1999).

38. Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115 (1998).

39. Hebsgaard, S. M. *et al.* Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucleic Acids Res.* **24,** 3430–3452 (1996).

40. Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. A tool for analyzing and annotating genomic sequences. *Genomics* **46,** 37–45 (1997).

41. Frishman, D. & Mewes, H.-W. PEDANTic genome analysis. *Trends Genet.* **13,** 415–416 (1997).

42. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).

43. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300,** 1005–1016 (2000).

Supplementary Information is available at *Nature*'s World-Wide Web site (http://www.nature.com) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank K. Mayer and H. Schoof of MIPS for discussions; S. Rhee and E. Huala of TAIR for sequences for the RI markers; and R. Wells for editing the manuscript. This work was funded by National Science Foundation/US Department of Energy/US Department of Agriculture (NSF/DOE/USDA) grants to the SPP Consortium and TIGR.

Correspondence and requests for materials should be addressed to A.T. (e-mail: theo@nature.berkeley.edu) or J.R.E. (e-mail: ecker@salk.edu). The accession members for chromosome 1 are: northern arm, AE005172; southern arm AE005173.

Sequence and analysis of chromosome 3 of the plant *Arabidopsis thaliana*

European Union Chromosome 3 Arabidopsis Sequencing Consortium, The Institute for Genomic Research & Kazusa DNA Research Institute*

* *A full list of authors appears at the end of this paper*

Arabidopsis thaliana **is an important model system for plant biologists[1]. In 1996 an international collaboration (the Arabidopsis Genome Initiative) was formed to sequence the whole genome of *Arabidopsis*[2] and in 1999 the sequence of the first two chromosomes was reported[3,4]. The sequence of the last three chromosomes and an analysis of the whole genome are reported in this issue[5–7]. Here we present the sequence of chromosome 3, organized into four sequence segments (contigs). The two largest (13.5 and 9.2 Mb) correspond to the top (long) and the bottom (short) arms of chromosome 3, and the two small contigs are located in the genetically defined centromere[8]. This chromosome encodes 5,220 of the roughly 25,500 predicted protein-coding genes in the genome. About 20% of the predicted proteins have significant homology to proteins in eukaryotic genomes for which the complete sequence is available, pointing to important conserved cellular functions among eukaryotes.**

Chromosome 3 is submetacentric and represents about 20% of the *Arabidopsis* genome. It has been estimated, using yeast artificial chromosome (YAC)-based physical maps[9,10], to be 21–23 Mb long (excluding the centromeric and telomeric regions). We sequenced 330 clones (bacterial artificial chromosomes (BACs)[11,12], P1 clones[13] and transformation-competent artificial chromosomes (TACs)[14]) and eight polymerase chain reaction (PCR) products and assembled them into four contigs representing 23,172,617 base pairs (bp) of non-redundant sequence. The bottom arm contains a residual sequencing gap of around 5 kilobases (kb). Of the approximately 150 and 450 kb of sequence in the two small centromeric contigs, 340 kb (90 and 250 kb, respectively) correspond to high accuracy DNA sequence; the rest consists of unfinished highly repetitive BAC sequences.

For each chromosome arm, the canonical telomeric repeats[15] specific for chromosome ends border a long euchromatic region (~11 and ~7 megabases (Mb) for the top and bottom arms, respectively) characterized by a high and roughly uniform gene density. The gene density then gradually decreases as the retrotransposon density increases towards the peri-centromeric/centromeric heterochromatic region. The top arm contig terminates at the F15D2 BAC clone, which contains at its end a 180-bp tandem repeat characteristic of the *Arabidopsis* centromere[16,17]. The bottom arm contig begins at the F4M19 BAC clone with a 5S ribosomal DNA (rDNA) repeat cluster[18]. The two small centromeric contigs were mapped in between the two arm contigs by tetrad analysis[8]. The relative positions and orientations of these small contigs have not yet been confirmed experimentally. The probable structure of the chromosome 3 centromere is shown in ref. 7 and in Supplementary Information. The size of the genetically defined centromeric region is estimated to be around 1.7 Mb; added to the size of the chromosome arms, this indicates a size of 24 Mb for the whole chromosome. Unexpectedly, the centromeric region contains, in addition to known repetitive elements[7], a block of 40 nearly perfect telomeric repeat units and a single complete rDNA unit (over 99% identical in the 25S, 18S and 5.8S regions). A general description of the characteristics of

Table 1 Features of chromosome 3

(a) The DNA molecule	
Length	23,172,617 bp
Top arm	13,590,268 bp
Bottom arm*	9,582,349 bp
Base composition (%GC)	
Overall	35.4
Coding	44.3
Non-coding	33.0
Number of genes	5,220
Gene density	4.5 kb per gene
Average gene length	1,925 bp
Average peptide length	424 amino acids
Exons	
Number	26,570
Total length	6,654,507 bp
Average per gene	5.1
Average size	250 bp
Introns	
Number	21,350
Total length	3,397,531 bp
Average size	159 bp
Percentage of genes with ESTs†	59.8%
Number of ESTs†	20,732
(b) The proteome	
Total proteins	5,220
Proteins with INTERPRO domains	2,989 (57.8%)
Genes which contain at least one transmembrane domain	1,615 (30.9%)
Genes which contain at least one SCOP domain	1,664 (31.9)%
Secretory pathway default value‡	877
Secretory pathway >0.95 specificity	813
Chloroplast default value	754
Chloroplast >0.95 specificity	420
Mitochondria default value	554
Mitochondria >0.95 specificity	63
Functional classification	
Cellular metabolism	745
Transcription	566
Plant defence	354
Signalling	356
Growth	357
Protein fate	314
Intracellular transport	269
Transport	155
Protein synthesis	148
Total	3,264

* The size of the bottom arm included the two small centromeric contigs (~340 kb).
† EST matches were calculated using a similarity threshold of 90%.
‡ The assignation to secretory pathway, chloroplast and mitochrondria result from a TargetP analysis.